Speeding up Browsing with Caching Regular Expressions

Renáta Hodován

It is clearly visible that the Web is becoming the most popular application platform. Static web pages fade into the past; more and more web sites use advanced techniques like server and client-side scripting, and AJAX to improve the user experience. For client-side scripting, JavaScript is by far the most commonly used dynamic language [6], it is supported by all major browsers. Not surprisingly, the optimization of JavaScript execution is a hot research topic [5,7]. However, as our experiments show, attention should be paid to regular expressions as well. In this work, we present a study on how regular expressions are used on popular web sites, introduce the idea of caching regular expressions, and present results of an experimental implementation.

First, the most popular web sites [1,2,4,8] have been collected in four categories (community, news, torrent, and adult) and then we visited those pages. With a modified version of WebKit [3] – a popular browser engine used in several desktop and mobile browsers – all regular expression patterns has been logged. Thus, we got a list of patterns representing the real-life load of a regular expression engine. Table 1 shows the number of regular expressions parsed and compiled during the browsing sessions, and the number of unique patterns as well. It is interesting to see how low the ratio of unique regular expressions is: less than 4%, irrespectively of the type of the visited pages.

Table 1: Regular expressions on web pages

	Community	News	Torrents	Adult
Total	50243	63165	22605	48666
Unique	1529(3.04%)	1972(3.12%)	889 (3.93%)	701 (1.43%)

The above results motivated us to experiment with caching. A cache has been implemented where the internal representations of the most recently compiled regular expressions are stored. Round-robin caching policy has been used with a cache size of 256. Table 2 shows the hit-miss ratio of the cache and Table 3 presents the effect of caching on the performance of regular expression parsing and compilation. (We experimented with the more complex Least Recently Used policy as well, but it was not significantly better.) As the latter Table shows, caching can result in a reduction in regular expression parsing and compilation time as high as 69%, which equals to a 3-fold speedup.

Table 2: Hit-miss ratio of the cache

	Community	News	Torrents	Adult
Hit	47240	28176	20575	46690
Miss	2963	3309	2030	1976

Table 3: The effect of caching on regex compilation time

	Community	News	Torrents	Adult
Original (ms)	354.862	190.309	170.395	197.131
Cached (ms)	124.826	73.417	52.088	86.809
Gain	65.19%	61.41%	69.43%	55.96%

References

[1] News Top 20. The pick of the best news sites on the net, May 2010.

[2] Inc Alexa Internet. Top sites by cathegory: Adult, May 2010. http://www.alexa.com/topsites/category/Top/Adult.

[3] Apple, Inc. TheWebKit open source project. http://webkit.org/ (Accessed 1 July 2009).

[4] eBizMBA. Top 15 most popular torrent websites, May 2010. http://www.ebizmba.com/articles/torrent-websites.

[5] Andreas Gal, Brendan Eich, Mike Shaver, David Anderson, David Mandelin, Mohammad R. Haghighat, Blake Kaplan, Graydon Hoare, Boris Zbarsky, Jason Orendorff, Jesse Ruderman, Edwin Smith, Rick Reitmaier, Michael Bebenita, Mason Chang, and Michael Franz. Trace-based just-in-time type specialization for dynamic languages: *In Proceedings of the 2009 ACMSIGPLAN conference on Programming language design and implementation (PLDI'09), pages 465-478, Dublin, Ireland, June 2009.*

[6] Zoltán Herczeg, Gábor Lóki, Tamás Szirbucz, and Ákos Kiss: Guidelines for JavaScript programs: Are they still necessary? *In Proceedings of the 11th Symposium on Programming Languages and Software Tools (SPLST'09) and 7th Nordic Workshop on Model Driven Software Engineering (NW-MODE'09), pages 59-71, Tampere, Finland, August 26-28, 2009. Tampere University of Technology.*

[7] Maciej Stachowiak: Introducing SquirrelFish Extreme, September 2008. http://webkit.org/blog/214/introducing-squirrelfish-extreme/ (Accessed 1 July 2009).

[8] Wikipedia. List of social networking websites, May 2010. http://en.wikipedia.org/wiki/List_of_social_networking_website.